Biomedical Informatics 231 Final Project
Yang Bi

A critical review on chromatin architecture modeling with high throughput chromosome conformation capture techniques

Introduction

Chromosomes are actively packed into dynamic three dimensional structures inside the cell nucleus. Mounting evidence suggests this organization is critical for many aspects of genome regulation such as gene expression, DNA replication, chromosome transmission and genome stability.[1-3] Our current understanding of the nuclear architecture is limited to the two extremes of resolution. On the chromosomal level, we depict each chromosome occupying a different nuclear region termed chromosomal territories (CT).[1] On the scale of several hundred base pairs, the chromatin is generally viewed as nucleosomes connected by linkers or the "beads on a string" model. [4-6] This huge resolution gap has to be filled before any unifying model on genome architecture could be attempted.

A number of methods have been employed in the past decades to study genome wide higher order chromatin structure, including DNase I, micrococcal nuclease and Sono-seq[7]. While useful to certain extent, these methods provide limited spatial information on chromatin arrangement.

Chromosome conformation capture (3C) is a technique to study long range chromatin interactions.[8] Its high throughput derivatives, such as Hi-C[9] and tethered conformation capture (TCC) [10], are particularly informative in deciphering genome wide chromatin contacts on the Megabase (Mb) or even 100 Kilobase (Kb) scales, making them promising innovations to close the resolution gap in our understanding of genome architecture.

This review will start with a brief description of the experimental procedures of 3C and 3C based high throughput methods and then focus on recent advancements in the application of high throughput 3C based methods. Special attention will be paid to efforts dedicated to improve the experimental procedures and modeling of three dimensional chromatin architecture.

Experimental procedures of 3C and 3C based methods.

The 3C experiment consists of five steps (Figure 1). The first step is to cross link the DNA and associated protein complexes with formaldehyde. This fixation gives a snapshot of the *in vivo* interaction between DNA and proteins. This is followed by digestion of the cross-linked chromatin complex with a restriction enzyme that cuts DNA in the region of interest. The third step is to ligate the restricted chromatin at low DNA concentration to facilitate intramolecular ligation. The chromatin complexes are then reverse cross-linked and the DNA is purified. If two distal chromatin regions are brought to proximity by chromatin looping, they are likely to be ligated. In the last step, purified DNA is then used as template for quantitative PCR reactions to determine the rate of ligation of distal regions of interest. If this rate of contact is higher than a control, usually contact rate of intervening regions as probability of contact decreases as distance increases, the two distal regions are likely to form a chromatin loop.[8]

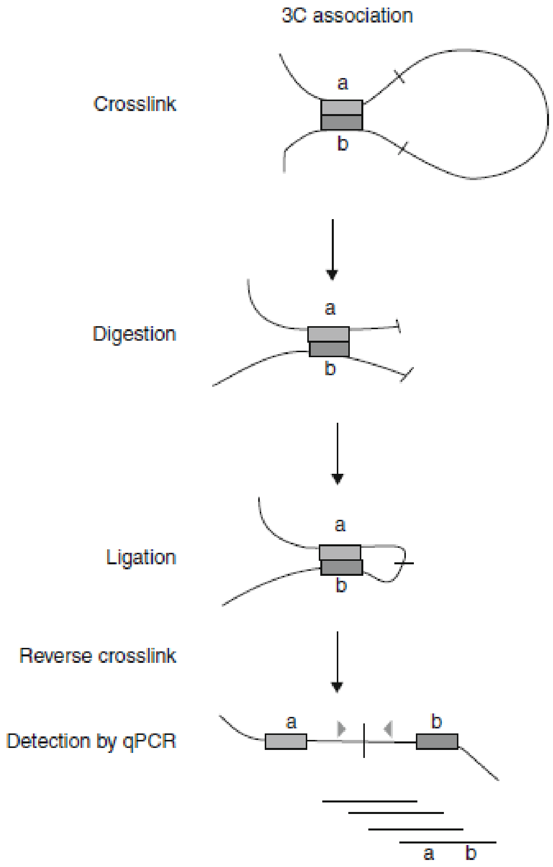

Fig. 1. 3C procedure. Schematic representation of a 3C assay. *Light grey* and *dark grey boxes* represent two interacting genomic elements, a and b, that are separated by a long intervening region (*black curved line*). The 3C assay starts with a cross-linking step using formaldehyde to capture protein–protein and protein–DNA interactions. A second step consists of enzymatic digestion with a restriction enzyme known to recognize sites in the investigated regions. In the third step, the cross-linked complex is religated under conditions that favor intramolecular ligation. Lastly, the cross-links are reversed by heat treatment; the DNA is then purified and the resulting 3C product is detected by qPCR. Adopted from Nativio et al. 2012.[11]

While useful for specific loci of interest, 3C has very limited throughput. Several modifications, including circularized chromosome conformation capture (4C) and carbon copy chromosome conformation capture (5C), have been made to increase its capacity (Figure 2). 4C allows study of contact between one region of interest to all other regions of the genome. It uses a second restriction enzyme after first ligation and then circularize the ligated product. This circularized DNA is used as template for PCR reactions to determine genomic regions that are ligated to the region of interest.[12] 5C is very similar to 3C but a set of multiplex primers connected to two universal primers are used for the PCR reaction. The PCR products are then sequenced to determined ligated sequences. Depending on the number of multiplex primers, this method establish interaction between many loci to many loci simultaneously.[13]

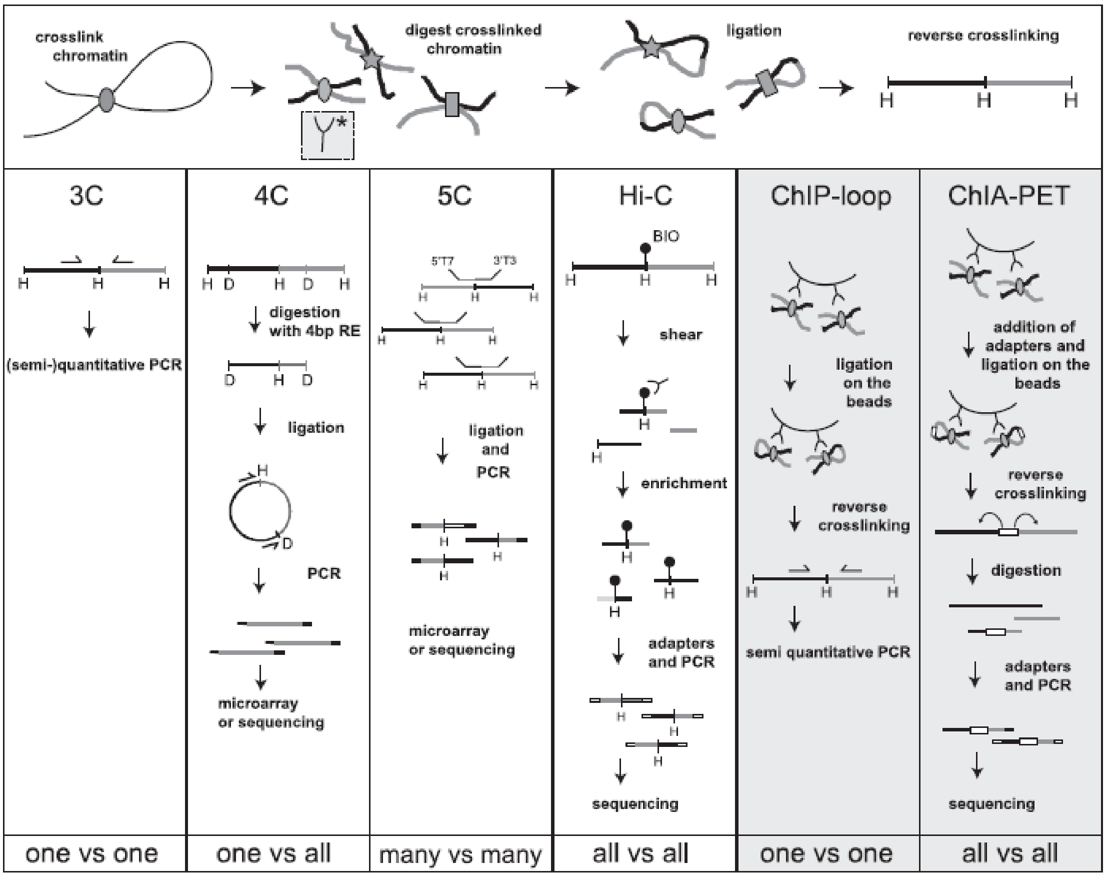

Figure 2. Overview of 3C-derived methods. An overview of the 3C-derived methods that are discussed is given. The horizontal panel shows the cross-linking, digestion, and ligation steps common to all of the "C" methods. The vertical panels indicate the steps that are specific to separate methods. Adopted from de Wit E et al. 2012.[14]

It is not until the development of Hi-C that study of chromatin interactions at the genome wide level becomes possible.[9,15] In general, this method involves digestion of the chromatin with a restriction enzyme and subsequently filling up 5' overhangs with nucleotides in which one type of nucleotide is biotinylated. Ligation is carried out after chromatin dilution to favor intramolecular ligation events. Ligated fragments with biotin at their ligation junctions are isolated by streptavidin and identified by high-throughput sequencing. Meaningful interpretation of this huge amount of data depends on effective and robust statistical analysis. The output paired end sequencing data is first mapped to the genome with established statistical programs.[16] A quality control of reads is usually carried out at this step to discard sequences that are unlikely to rise from ligation of restricted fragments if they are too far away from the nearest restriction sites.[17] The core analytic step is to create a matrix comparing the observed number of reads between two loci of certain length, or more commonly referred to as bins, to the expected number of reads between these two bins (Figure 3). The observed interaction matrix is a collection of the numbers of interactions between each pair of bins from the mapped data. The expected number of interactions between two intrachromosomal bins is derived from the experimental data by taking the total number of observed interactions at a distance s divided by the total number of possible interactions at distance *s* across all chromosomes. To obtain the interchromosomal averages, the number of observed interactions between bins on a pair of chromosomes was divided by the number of possible interactions between the two chromosomes. The observed interaction matrix can then be normalized by the expected interaction matrix to generate an Observed/Expected matrix. Because two bins that are close together in space should interact with similar bins and thus should have correlated interaction profiles, a correlation matrix can be derived by calculating the Spearman correlation coefficient (Pearson correlation assumes linearity) between each pair of bins (or each pair of vectors in the Observed/Expected matrix). This correlation matrix reveals the average positioning of bins with respects with each others.[15]

The high throughput methods can be easily coupled with chromatin immunoprecipitation (ChIP) to build a chromatin interactome map of loci bound by a specific protein of interest. The major modification is that ChIP is carried out after crosslinking and sonification to enrich for fragments bound by a particular protein of interest. [18,19] One such method, known as ChIP-PET, effectively reveal genome wide chromatin interactions associated with a protein of interest,[18,19] allowing interpretations of the functional involvement of the protein of interest in modifying chromatin structure.

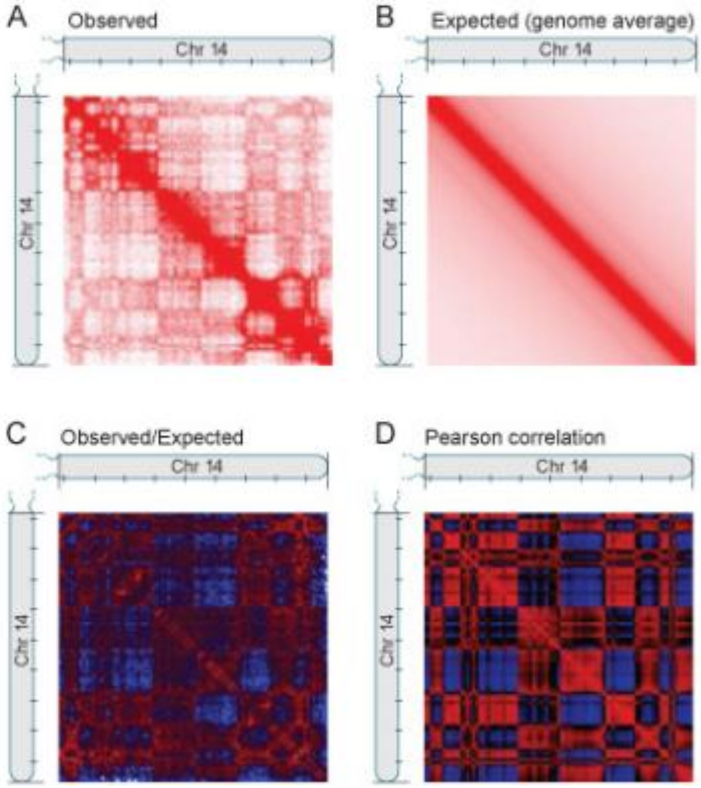

Figure 3. Heat maps depicting intrachromosomal contact heat maps for chromosome 14 at resolution of 1Mb. A) Observed interactions. B) Expected interaction frequencies based on genomic distance. C) Quotient of matrices A and B, showing more (red) or less (blue) interactions than expected. D) Correlation matrix between intrachromosomal interaction profiles. Adopted from Lieberman-Aiden, E. et al. 2009.

Recent developments on 3C based high throughput methods

3C based high throughput methods gained popularity shortly after the first Hi-C paper. Genome-wide contact maps have been obtained for Lymphoblastoid cells[10,15], mouse[20,21], fission yeast[22], budding yeast[9,23] and fruit fly[24]. These studies have demonstrated the potential of Hi-C methods in probing and explaining various biological processes. For instance, spatial proximity contributes to translocations via formation of double strand breaks[20]; clusters of interacting loci isolated from each other by insulators are identified in mouse genome[21]; actively transcribed genes, genes with similar promoter elements (under the regulation of similar groups of proteins) and functionally related genes tend to associate with each other[22,24]. However, given the complicated nature of the experiments, 3C based methods are prone to bias and misinterpretations. A number of studies focus on the improvement of experimental designs and quality control of experimental data. Also, there is an emerging effort to construct three dimensional models from high throughput 3C based methods.

Modifications in experimental procedures

A major source of error in 3C based methods comes from random intermolecular collision of restricted DNA fragments that are not cross-linked to each other. Because randomly selected DNA fragments are more likely to originate from different chromosomes, these ligations tend to be overwhelmingly interchromosomal.[10] Such random noise would mask a substantial number of interchromosomal interactions, most of which occur at low frequencies. In addition, sample cells used for the experiments represent a heterogeneous pool of chromatin architecture and many interesting and informative chromatin loopings are only present in a fraction of the cells.[25,26] These low frequency interactions are also likely to be lost due to the high noise to signal ratio.

One way to tackle this problem is to carry out multiple control experiments on uncrosslinked purified genomic DNA to establish a random intermolecular interaction background and compare data from actual experiments with the background to rid of the noise. However, this method is both tedious and expensive and is not shown to be more effective than normalization with an expected interaction matrix.[9]

Kalhor R et al. took an alternative approach by carrying key steps of the experiments on solid phase instead of solutions. Proteins are biotinylated prior to DNA digestion and protein bound fragments are immobilized at a low surface density on streptavidin-coated beads. Ligation is carried out while the DNA fragments are immobilized on the beads, hence the rate of background ligation due to random collisions between molecules is largely reduced.

This innovation, termed tether conformation capture (TCC), greatly reduces the noise to signal ratio and unravels informative details on interchromosomal interactions missed by traditional Hi-C methods (Figure 4). The investigators first demonstrated that the TCC method could reproduce the intrachromosomal interaction pattern generated by Hi-C method (Figure 4a 4b). They tested the effectiveness of their modification with two restriction enzymes. Using the traditional 6 cutter (recognize 6 base pairs) HindIII, the TCC method generated only half the interchromosomal interactions generated by Hi-C (Figure 4c). Using a 4 cutter restriction enzyme MboI, the TCC method generated less than half the interchromosomal interaction generated by Hi-C. Interestingly, the TCC method generated a comparable level of interchromosomal interactions with 4 cutter restriction enzyme to that generated by 6 cutter. 4 cutter restriction enzymes generate more but smaller DNA fragments which have higher random collision probability than that generated by 6 cutters. This method demonstrated the possibility of using 4 cutter restriction enzymes without greatly compromising signal to noise ratio. As 4 cutter enzymes cut much more frequently, this technique would allow study of chromatin architecture at a much higher resolution.

Reducing random interchromosomal interaction decreases the expected interchromosomal interactions and thus increases the Observed/Expected value. As expected, the TCC method captured a lot of interchromosomal interactions otherwise missed by Hi-C method (Figure 4d 4e).

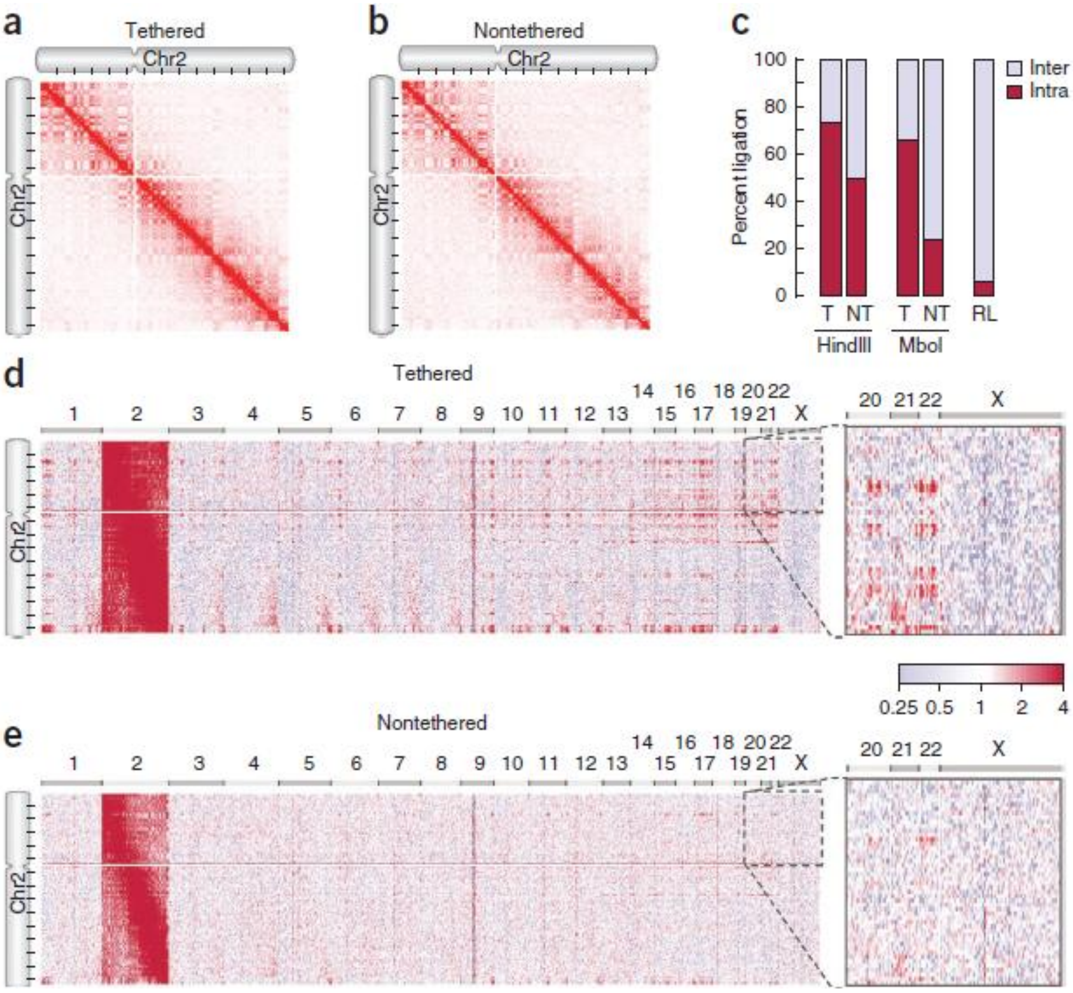

Figure 4. Tethering improves the signal-to-noise ratio of conformation capture. (**a**,**b**) TCC can reproduce the results obtained by Hi-C10. A genome-wide contact frequency map is compiled from the ligation frequency data generated by tethered (TCC) (**a**) and nontethered (Hi-C) (**b**) conformation capture. The portion of each map that corresponds to the intrachromosomal contacts of chromosome 2 is shown. The intensity of the red color in each position of the map represents the observed frequency of contact between corresponding segments of the chromosome (**c**) The observed fractions of intra- and interchromosomal ligations in tethered (T) and nontethered (NT) libraries produced using HindIII or MboI. The random ligation (RL) bar represents the expected fractions if all ligations occurred between noncrosslinked DNA fragments. For the nontethered MboI library only, these fractions were determined by sequencing 160 individual DNA molecules from three replicates of the experiment. (**d**,**e**) The genome-wide enrichment map for chromosome 2, compiled from the tethered (**d**) and nontethered (**e**) HindIII libraries. Enrichment is calculated as the ratio of the observed frequency in each position to its expected value; expected values were obtained assuming completely random ligations (Online Methods). Red and light blue, respectively, indicate enrichment and depletion of a contact. Chromosome 2 (left) extends along the *y* axis whereas all 23 chromosomes (top) extend along the *x* axis. Adopted from Kalhor R et al. 2011

Generation of robust interaction frequency matrices

The correlation matrix is derived from the quotient matrix of Observed interaction matrix/Expected interaction matrix and is used as a measure of spatial relationship between loci. Generation of robust observed interaction frequency matrix and/or expected interaction frequency matrix is therefore critical for drawing any reliable conclusion on the chromatin structure. Pioneering Hi-C studies make use of the idealized expected number of contacts between each pair of bins to construct the Expected matrix without considering potential bias and artifacts introduced by the experimental procedures.

Five major sources of bias were identified in two studies: the length of restriction fragments, the length of fragment ends, GC content of the paired-end reads, the length of DNA segments at the circularization steps and "mappability" of sequence reads. [27,28]

In theory, as the number of positions accessible to fixating along a restriction fragment increases with its size, the interaction probability increases linearly with restriction fragment size. Indeed, this is true for restriction fragments under 800 bp. However, for longer restriction fragments, a plateau is reached, suggesting that the maximum probability for at least one cross-linking event to occur along that length is reached (Figure 5).

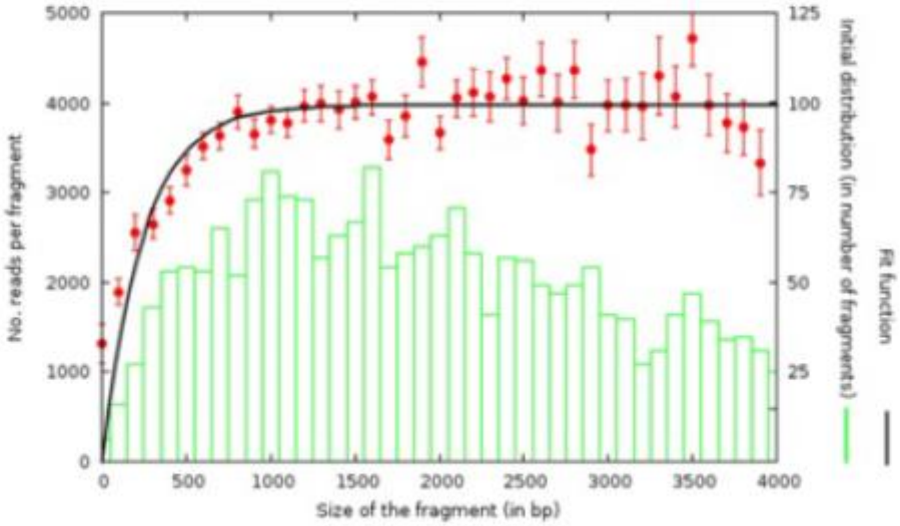

Fig 5. Quantification of the fragment length bias. Adopted from Cournac, A. et al. 2012.

The sizes of crosslinked fragment ends affect their ligation efficiency. Ligation efficiency is low when fragment ends are too long or too short. Optimal ligation takes place when both crosslinked fragment ends have intermediate length (Figure 6).

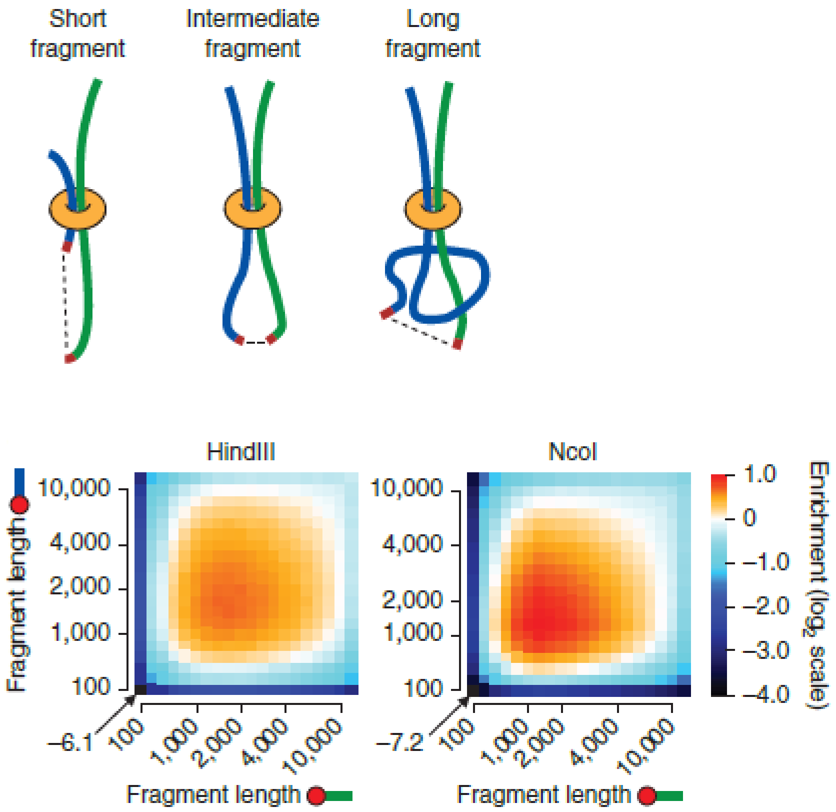

Fig 6. Schematic drawing and quantification of ligation efficiency due different fragment length. The trans Hi-C coverage enrichment is defined as the ratio between the observed number of trans contacts and the total number of assayed fragment pairs. Adopted from Yaffe, E. et al. 2011.

Another potential bias comes from the GC content of the restriction fragments. Fragments with extreme GC content are underrepresented in the final interaction reads. Deep sequencing appears to favor reads with a GC content of about 45% (Figure 7).

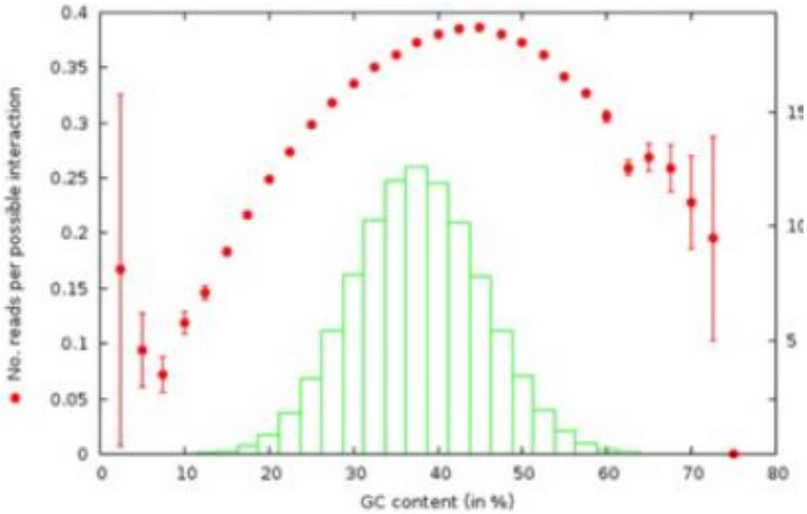

Fig 7. Quantification of the fragment length bias. Adopted from Cournac, A. et al. 2012.

The forth bias is dependent on the length of DNA segments at the circularization steps (Most human Hi-C protocol does not require a circularization step). Mechanical property of the DNA polymer dictates too small a fragment will be poorly ligated due to high bending persistence while too long a fragment will also disfavor ligation due to entropic contribution to the free energy. Optimal circularization is achieved at around 500 bp (Figure 8). In addition, this bias is highly non-monotonous in cycles of 10.5 bp. For instance, it favors the circularization length of 261 bp, but circularization length of disfavor 266 bp and again favor circularization length of 271 bp.

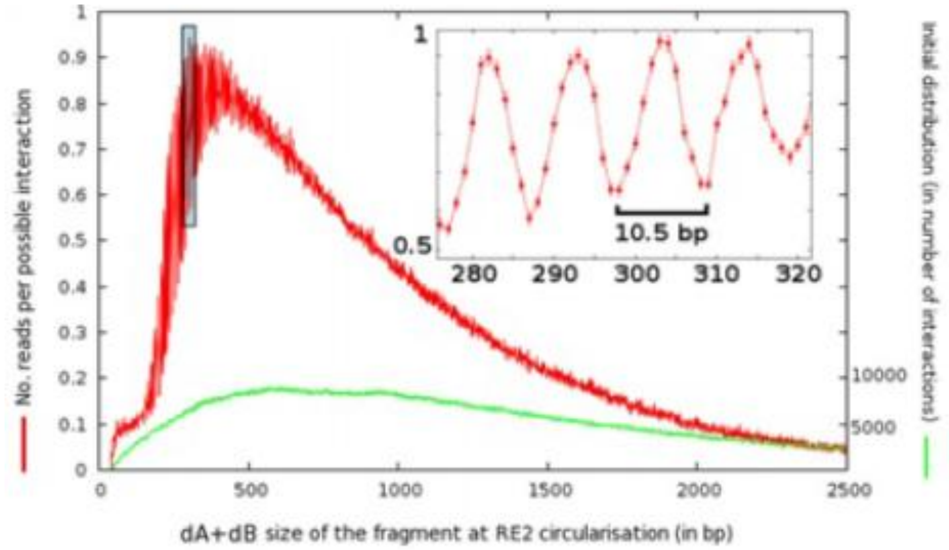

Fig 8. Quantification of the circularization length bias. Adopted from Cournac, A. et al. 2012.

The last bias is associated with highly repetitive regions. Regions with low level of unique sequences are usually unmappable and hence underrepresented in the final interaction reads (Figure 9).

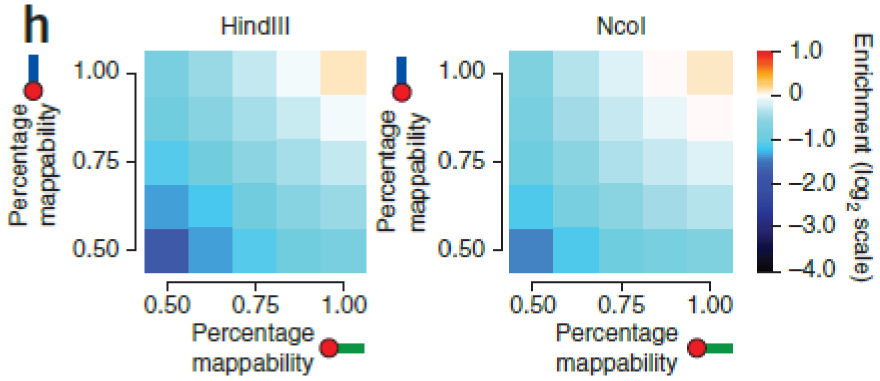

Fig 9. Quantification of the mappability bias. Adopted from Yaffe, E. et al. 2011.

As bins used for later computation each contains one or more fragment ends, the bias is carried on to later computations. To eliminate these systematic biases, Yaffe, E et al. 2011 developed a probabilistic model to generate a more accurate expected interaction matrix *in silico*.[28] This method was designed for human Hi-C data analysis and correct for fragment ends length, GC content and mappability.

Fragment ends from chimeric Hi-C reads were binned according to the length of their corresponding fragments into 20 equally sized bins denoted by $(\mathrm{B}_{\mathrm{i}}^{\mathrm{len}})_{\mathrm{i}=1}^{20}$. The seed matrix for fragment lengths is defined as : $\mathrm{S}_{\mathrm{len}}[\mathrm{i},\mathrm{j}] = (1/\mathrm{P}_{\mathrm{prior}}) * \frac{\mathrm{O}_{\mathrm{len}}[\mathrm{i},\mathrm{j}]}{\mathrm{T}_{\mathrm{len}}[\mathrm{i},\mathrm{j}]}$, where $\mathrm{P}_{\mathrm{prior}}$ is the prior probability to observe a pair and is equal to the total number of observed interchromosomal pairs divided by the total number of possible interchromosomal pairs, $\mathrm{O}_{\mathrm{len}}[\mathrm{i},\mathrm{j}]$ is the number of observed interchromosomal pairs such that one fragment end is in bin $\mathrm{B}_{\mathrm{i}}^{\mathrm{len}}$ and the other is in bin $\mathrm{B}_{\mathrm{j}}^{\mathrm{len}}$, and $\mathrm{T}_{\mathrm{len}}[\mathrm{i},\mathrm{j}]$ is the total number of possible unique interchromosomal pairs such that one fragment end is in bin $\mathrm{B}_{\mathrm{i}}^{\mathrm{len}}$ and the other is in bin $\mathrm{B}_{\mathrm{j}}^{\mathrm{len}}$. The function is more amenable to understanding in the form of $\mathrm{S}_{\mathrm{len}}[\mathrm{i},\mathrm{j}] = \frac{\mathrm{O}_{\mathrm{len}}[\mathrm{i},\mathrm{j}]}{\mathrm{P}_{\mathrm{prior}} * \mathrm{T}_{\mathrm{len}}[\mathrm{i},\mathrm{j}]}$, which simply represents the ratio between the number of observed interchromosomal pairs such that one fragment end is in bin $\mathrm{B}_{\mathrm{i}}^{\mathrm{len}}$ and the other is in bin $\mathrm{B}_{\mathrm{j}}^{\mathrm{len}}$ and the number of such possible unique interchromosomal pairs that are expected to be observed on average.

The GC content seed matrix S_{gc} is computed in a similar manner where 20 bins are defined according to the GC content of the 200 bp near the fragment end. Same procedure is applied to mappability matrix but with only five bins (not a seed matrix because mappability scores can be determined from the empirical data).

The expected interaction probability for two given fragment ends a and b is defined as: $\mathrm{P}(\mathrm{X}_{\mathrm{a,b}}) = \mathrm{P}_{\mathrm{prior}} * \mathrm{F}_{\mathrm{len}}(\mathrm{a}_{\mathrm{len}}, \mathrm{b}_{\mathrm{len}}) * \mathrm{F}_{\mathrm{gc}}(\mathrm{a}_{\mathrm{gc}}, \mathrm{b}_{\mathrm{gc}}) * \mathrm{M}(\mathrm{a}) * \mathrm{M}(\mathrm{b}),$ where $\mathrm{a}_{\mathrm{len}}, \mathrm{b}_{\mathrm{len}},$ a_{gc} and b_{gc} are the fragment length bins and GC content bins of the two ends, $\mathrm{M}(\mathrm{a})$ and $\mathrm{M}(\mathrm{b})$ are mappability scores of the two ends, and $\mathrm{F}_{\mathrm{len}}(\mathrm{a}_{\mathrm{len}}, \mathrm{b}_{\mathrm{len}})$ and $\mathrm{F}_{\mathrm{gc}}(\mathrm{a}_{\mathrm{gc}}, \mathrm{b}_{\mathrm{gc}})$ are two real valued functions. Mappability scores are readily calculated from the empirical data while $\mathrm{F}_{\mathrm{len}}(\mathrm{a}_{\mathrm{len}}, \mathrm{b}_{\mathrm{len}})$ and $\mathrm{F}_{\mathrm{gc}}(\mathrm{a}_{\mathrm{gc}}, \mathrm{b}_{\mathrm{gc}})$ are determined statistically through a maximum likelihood method. The likelihood function is:

$$\mathrm{L}(\mathrm{F}_{\mathrm{len}}, \mathrm{F}_{\mathrm{gc}}) = \prod_{\{\mathrm{a,b}\}\in \mathrm{I}} \mathrm{P}(\mathrm{X}_{\mathrm{a,b}}) * \prod_{\{\mathrm{a,b}\}\notin \mathrm{I}} \left(1 - \mathrm{P}(\mathrm{X}_{\mathrm{a,b}})\right)$$

$$= \prod_{\mathrm{c}=(\mathrm{a}_{\mathrm{len},}\ \mathrm{a}_{\mathrm{gc},}\ \mathrm{b}_{\mathrm{len},}\ \mathrm{b}_{\mathrm{gc}}\ )} \mathrm{P}(\mathrm{X}_{\mathrm{a,b}})^{\mathrm{n}_{\mathrm{c}}} * [1 - \mathrm{P}(\mathrm{X}_{\mathrm{a,b}})]^{\mathrm{m}_{\mathrm{c}}}$$

where I is the set of observed fragment end pairs, n_{c} is the number of observed pairs that match the bin criteria of c and m_{c} is the number of observed pairs match the criteria but are not observed. Solving the equation gives the values of $\mathrm{F}_{\mathrm{len}}(\mathrm{a}_{\mathrm{len}}, \mathrm{b}_{\mathrm{len}})$ and $\mathrm{F}_{\mathrm{gc}}(\mathrm{a}_{\mathrm{gc}}, \mathrm{b}_{\mathrm{gc}})$ that maximize the probability of observe interaction matrix to occur. However, this equation can only be solved through heuristic method. Here the two seed matrices are used as initial input, $\mathrm{F}_{\mathrm{len}}^{0} = \mathrm{S}_{\mathrm{len}}$ and $\mathrm{F}_{\mathrm{gc}}^{0} = \mathrm{S}_{\mathrm{gc}}$. The likelihood function is maximized by alternating between the optimization of the two matrices:

$$\mathrm{F}_{\mathrm{len}}^{\mathrm{n+1}} = \underset{\mathrm{F}_{\mathrm{len}}}{\arg\max} \mathrm{L}(\mathrm{F}_{\mathrm{len}}, \mathrm{F}_{\mathrm{gc}}^{\mathrm{n}}), \mathrm{F}_{\mathrm{gc}}^{\mathrm{n+1}} = \mathrm{F}_{\mathrm{gc}}^{\mathrm{n}}$$

$$\mathrm{F}_{\mathrm{gc}}^{\mathrm{n+1}} = \underset{\mathrm{F}_{\mathrm{gc}}}{\arg\max} \mathrm{L}(\mathrm{F}_{\mathrm{len}}^{\mathrm{n}}, \mathrm{F}_{\mathrm{gc}}), \mathrm{F}_{\mathrm{len}}^{\mathrm{n+1}} = \mathrm{F}_{\mathrm{len}}^{\mathrm{n}}$$

This two steps are repeated until the improvement in the log-likelihood is smaller than an arbitrary threshold. Applying the resultant $\mathrm{F}_{\mathrm{len}}$ and F_{gc} matrices would generate an improved expected interaction matrix.

As current sequencing depth may yield less than 1 Hi-C read per bin, the investigator also smoothed the observed and expected contact matrices using linear weights:

$$\mathrm{O}_{[\mathrm{i,j}]}^{\mathrm{gw}} = \sum_{-\mathrm{W}<l<W,-W<k<W} \mathrm{O}_{[\mathrm{i+k,j+l}]} * \mathrm{w}_{\mathrm{k,l}},$$

where W=10 and $\mathrm{w}_{\mathrm{k,l}} = \frac{1}{|\mathrm{k}|+|\mathrm{l}|+1}$.

The same procedure is applied to expected interaction matrix and the quotient matrix is still calculated by Observed/Expected but further normalized by total coverage for different bin pairs.

This method is computationally expensive and only accounts for known sources of bias. An iterative correction method is later proposed to correct the observed interaction probability matrix instead of the expected interaction matrix. The method is less computationally demanding and does not rely on prior knowledge of the sources of bias. [29] It is built on the assumption that the bias for detecting interaction between two fragment ends is factorizable. The assumption is justified by showing that the method can explain 99.99% of the variance captured by Yaffe and Tanay's method.[29] The observed contact probability between two bins, $\mathrm{O_{ij}}$, is a realization of the expected observed contact probability, $\varepsilon_{\mathrm{ij}}$, with a certain distribution, $\mathrm{O_{ij}} \sim \mathrm{f}(\varepsilon_{\mathrm{ij}})$ (eg. Poisson distribution). The expected observed contact probability is defined as $\varepsilon_{\mathrm{ij}} = \mathrm{B_i} * \mathrm{B_j} * \mathrm{T_{ij}}$, where $\mathrm{B_i}$ and $\mathrm{B_j}$ are the biases at fragment end i and j and $\mathrm{T_{ij}}$ is the true matrix of interaction probability. Given the distribution f(.), the likelihood of the observed interaction probability matrix is given by:

$$\mathrm{L} = \prod_{\mathrm{ij}} \mathrm{f}(\mathrm{O_{ij}}, \mathrm{T_{ij}} \mathrm{B_i}\, \mathrm{B_j})$$

also, $\sum_{\mathrm{i}} \mathrm{T_{ij}} = 1,$ again taking the maximum likelihood approach, the equation can be solved to find the vector $\mathrm{B_i}$ which maximizes the likelihood of the observed interaction probability matrix and hence solve for $\mathrm{T_{ij}}$, the true interaction probability matrix. Assuming a Poisson distribution and setting first and second derivatives to 0, the maximum likelihood equation can be simplified to $\sum_{\mathrm{i}} \frac{\mathrm{O_{ij}}}{\mathrm{B_i B_j}} = 1$, which is solved by iterations.

While these two methods are generally accepted and applied in more recent studies[20,21,24], I would like to raise the caveat that high correlation between two vectors in the interaction probability matrix should not be simply interpreted as spatial proximity of two loci, but rather a sum of spatial proximity and behavioral similarity. It is likely that a large number of chromatin loops are highly flexible and interact with multiple loci with certain probability. When a snap shot of the sum of such probabilistic event is taken ,each locus should be seen as moving within a dynamic range with certain probability. The level of correlation of the interaction profiles of two loci (two vectors on the interaction matrix) depends on both spatial proximity and similarity of their dynamic movements. This way of interpreting the interaction frequency matrix has profound implications on building three dimensional models from the experimental data.

Interpretation of the 3C based high throughput data: towards the building of a 3D chromatin distance model on the genome scale

Despite our increasing capability to generate huge data set and to conduct robust data normalization, constructing a three dimensional distance model on the genome scale remains a challenging task due to the lack of suitable statistical methods.

The traditional method was to convert the question at hand to a constrained optimization problem.[30] Duan et al. 2010 made the first attempt to build a genome wide three dimensional model in yeast with this method.[9] Chromosomes are represented as series of beads in 3D space, spaced 10 kb apart. Each restriction fragment is mapped to its closest beads and an algorithm is carried out to place each beads at a distance that is inversely proportional to their interaction frequency (130 bp is assigned a length of 1 nm). A number of restrictions are introduced on the location of these beads to ensure a biologically possible spatial distribution: all beads must lie within a spherical nucleus with radius 1 μm; the distance between every two beads adjacent on the chromosome must be within a given range, 66nm to 99nm (estimated length for 10kb); no two beads should be placed closer than 30 nm (the thickness of chromatin); beads of different chromosomes must be separated by a minimum distance of 75nm to prevent interchromosomal crossings between segments connecting adjacent beads; rDNA is constrained to a spherical nucleolus with radius 0.3 μm; centromeres are placed on the diametrically opposite side of the nucleolus. This algorithm turns the problem into a nonlinear constrained optimization problem which is solved using an open-source software IPOPT.

This approach has two drawbacks. Firstly, the objective function (root mean square deviation) assumes that each IF measurement is equally reliable. Secondly, the structure obtained has no measure of uncertainty. A heuristic approach was then proposed to generate sets of candidate structures.[31,32]

Arguing these approaches are not based on probabilistic models and hence may not produce structures representative of the true set of possible structures, Rousseau et al. 2011 developed a probabilistic model using a Markov chain Monte Carlo-based method named MCMC5C.[33]Their algorithm samples from the posterior distribution of spatial positions of fragment ends given the observed interaction frequency. An ensemble of conformations is produced with probability equal to its posterior probability. The method starts with the generation of list of possible structures. A random structure $\mathrm{R}_0$is initially chosen to seed the selection process (t=0). A random perturbation is applied to R_t to generate R'_t. If R'_t has bigger posterior probability than R_t, it is retained and set to $\mathrm{R}_{\mathrm{t}+1}$. Otherwise, R'_t is retained with probability equal to the posterior probability of R'_t divided by the posterior probability of R_t. Posterior probability is calculated assuming Hi-C read counts follow a binomial probability distribution (a Gaussian distribution is used to approximate the binomial distribution for easier computation). The process is repeated to generate a list of structures $\mathrm{R}_1, \dots, \mathrm{R}_\mathrm{k}$. At some point m, for all k≥m, R_k becomes independent of R_0. In fact ,for any sufficiently large σ, $\mathrm{R}_{\mathrm{k}+\sigma}$ is independent of R_k. R_k, $\mathrm{R}_{\mathrm{k}+\sigma}$, $\mathrm{R}_{\mathrm{k}+2\sigma}$... $\mathrm{R}_{\mathrm{k}+(\mathrm{N}-1)\sigma}$are collected and named X_1 to X_N. This is the set of structure that is representative of the distribution of structures that fit the observed interaction frequency data. The structures are cluster hierarchically according to their level of similarity measured by root mean squared deviation. This clustering is used to determine the existence and number of structure subfamilies and the members of each subfamily.

The major problem with the method is that the Gaussian variance of each read count of a pair of loci cannot be accurately estimated since a single Hi-C interaction matrix does not contain enough information. Most recently, another method base on Bayesian inference is published.[34] The investigators believe that each topological domain on the chromatin share a consensus 3D chromosomal structure to keep its conservative functional forms. Based on this assumption, they developed a algorithm called Bayesian 3D Constructor for Hi-C data (BACH) to build consensus 3D structures for individual topological domains. They also believe adjacent topological domains exhibit flexibility and interact with each other in ways similar to that between loci and develop BACH-MIX to model it. For the BACH algorithm, they assume that the off-diagonal count $\mu_{ij}(i\neq j)$ in the interaction matrix follows a Poisson distribution with a rate θ_{ij}, where:

$\log(\theta_{ij})=\beta_0+\beta_1 \log(d_{ij}) + \beta_{enz} \log(e_i e_j) + \beta_{gcc} \log(g_i g_j) + \beta_{map} \log(m_i m_j)$.

β_1 measures the magnitude of negative association between μ_{ij} and d_{ij} (distance between the two loci) and β_{enz}, β_{gcc} and β_{map} are coefficients for the fragment end length effect, GC content effect and mappability effect. Let $P=(P_1,\dots P_n)^T$ represent the Cartesian coordinates of the n loci of interest and let $\beta=(\beta_0, \beta_1, \beta_{enz}, \beta_{gcc}, \beta_{map})$ be the collection of all nuisance parameters. The joint likelihood for n loci of interest is of the form:

$$P(U|P,\beta) = \prod_{1\leq i\leq j\leq n} P(u_{ij}\,|\theta_{ij}) = \prod_{i\leq i\leq j\leq n} \frac{e^{-\theta_{ij}}\theta_{ij}^{u_{ij}}}{u_{ij}!}$$

The algorithm adopts a fully Bayesian approach with non-informative priors to establish the posterior probability. As a result, a large number of parameters needs to be estimated. It starts by obtaining initial values for nuisance parameters using Poisson regression approach. In the next step initial 3D chromosomal structure is obtained by sequential importance sampling. This is followed by a Gibbs sampler to refine the 3D chromosomal structure and nuisance parameters. The BACH-MIX algorithm adopt a similar procedure for adjacent topological domains, treating each domain as a loci to establish their relative positions. Due to the large number of parameters estimated and multiple heuristic sampling processes involved, the reliability of this method requires further clarification. Also, the assumption of a consensus local 3D structure may not hold as there are known cases of local flexibility in chromatin interaction.

The three methods described here certainly move us towards the final goal of building meaningful 3D chromatin distance model on the genome scale. For their respective drawbacks, however, I would hesitate to recognize any of these as potentially standard algorithms for such a purpose. The difficulty of building 3D chromatin distance model from high throughput data is at least three fold. On the biological level, the chromatin structure of even the same cell type at the same cycle is likely to be highly flexible locally and any model that fails to appreciate such dynamic will not be able to generate a biologically meaningful model. On the experimental level, a heterogeneous pool of cells is usually used, further complicating interpretation of the result to generate one unique 3D structure. On the algorithm level, the data is usually insufficient to estimate the huge number of parameters in a model with high confidence.

These problems may be partially solved by devising more dynamic 3D model, using single cell technique to achieve homogeneity in starting material and conducting multiple sets of experiments for parameter estimation. More importantly, integrating polymer physics into our 3D chromatin structure model could be critical.[35-37] Notably, a recent "strings and binders switch" model, which combines the features of a random walk polyer model and the effects of interactions mediated by diffusible factors, reproduces many of the biological properties of chromatin structures.[37] An encouraging analogy would be the contribution of polymer physics in study of protein structure.[38]

Conclusion

This short review summarizes the experimental procedures of 3C and its high throughput derivatives and describes the recent advancements in experimental procedures, normalization of interaction frequency matrix and construction of 3D chromatin structures. Improvements and drawbacks on these methods and algorithms are discussed with some analysis of the difficulties in data analysis and 3D chromatin structure construction.

Studies have revealed interesting relationships between transcription factors, non-coding RNAs, cytoskeleton and the chromatin architecture.[18,19,39-41] Integrating all these information in the hope of building a unifying model of genome regulation requires the construction of three dimensional chromatin structure. Despite a number of difficulties, 3C based high throughput methods are arguably the most promising techniques for this purpose and they are constantly evolving.

Overcoming the hurdles in constructing a genome wide three dimensional chromatin architecture requires further understanding of fundamental properties of chromatin folding. For instance, more 3C studies should be conducted with deep sequencing to further our understanding of local chromatin dynamics. In addition, more studies of the polymer behavior of chromatin need to be carried out to provide insight into its physical properties. An interventionist approach can also be taken to manipulate linear and 3D chromatin to elucidate their properties: editing genomic features, adding artificial DNA-binding proteins or inserting reporter constructs and study how chromatin interactions are changed by perturbing certain factors.[42,43] These studies would be essential to improve the analysis of 3C based high throughput data and devising better models and algorithms in the construction of 3D chromatin models.

Reference.

1 Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**, 292-301, doi:10.1038/35066075
35066075 [pii] (2001).

2 Miele, A. & Dekker, J. Long-range chromosomal interactions and gene regulation. *Mol Biosyst* **4**, 1046-1057, doi:10.1039/b803580f (2008).

3 Misteli, T. & Soutoglou, E. The emerging role of nuclear architecture in DNA repair and genome maintenance. *Nat Rev Mol Cell Biol* **10**, 243-254, doi:10.1038/nrm2651
nrm2651 [pii] (2009).

4 Palmer, E. L., Gewiess, A., Harp, J. M., York, M. H. & Bunick, G. J. Large-scale production of palindrome DNA fragments. *Anal Biochem* **231**, 109-114, doi:S0003-2697(85)71509-6 [pii]
10.1006/abio.1995.1509 (1995).

5 Harp, J. M. *et al.* Preparative separation of nucleosome core particles containing defined-sequence DNA in multiple translational phases. *Electrophoresis* **16**, 1861-1864 (1995).

6 Harp, J. M. *et al.* X-ray diffraction analysis of crystals containing twofold symmetric nucleosome core particles. *Acta Crystallogr D Biol Crystallogr* **52**, 283-288, doi:10.1107/S0907444995009139
S0907444995009139 [pii] (1996).

7 Sajan, S. A. & Hawkins, R. D. Methods for identifying higher-order chromatin structure. *Annu Rev Genomics Hum Genet* **13**, 59-82, doi:10.1146/annurev-genom-090711-163818 (2012).

8 Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-1311, doi:10.1126/science.1067799
295/5558/1306 [pii] (2002).

9 Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465**, 363-367, doi:10.1038/nature08973
nature08973 [pii] (2010).

10 Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol* **30**, 90-98, doi:10.1038/nbt.2057
nbt.2057 [pii] (2012).

11 Nativio, R., Ito, Y. & Murrell, A. Quantitative chromosome conformation capture. *Methods Mol Biol* **925**, 173-185, doi:10.1007/978-1-62703-011-3_11 (2012).

12 Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* **38**, 1348-1354, doi:ng1896 [pii]
10.1038/ng1896 (2006).

13 Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* **16**, 1299-1309, doi:gr.5571506 [pii]
10.1101/gr.5571506 (2006).

14 de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev* **26**, 11-24, doi:10.1101/gad.179804.111

26/1/11 [pii] (2012).

15 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369

326/5950/289 [pii] (2009).

16 Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-1858, doi:10.1101/gr.078212.108

gr.078212.108 [pii] (2008).

17 van Berkum, N. L. *et al.* Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp*, doi:10.3791/1869

1869 [pii] (2010).

18 Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58-64, doi:Doi 10.1038/Nature08497 (2009).

19 Handoko, L. *et al.* CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* **43**, 630-638, doi:10.1038/ng.857

ng.857 [pii] (2011).

20 Zhang, Y. *et al.* Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148**, 908-921, doi:10.1016/j.cell.2012.02.002

S0092-8674(12)00158-4 [pii] (2012).

21 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082

nature11082 [pii] (2012).

22 Tanizawa, H. *et al.* Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res* **38**, 8164-8177, doi:10.1093/nar/gkq955

gkq955 [pii] (2010).

23 Rodley, C. D. M., Bertels, F., Jones, B. & O'Sullivan, J. M. Global identification of yeast chromosome interactions using Genome conformation capture. *Fungal Genet Biol* **46**, 879-886, doi:DOI 10.1016/j.fgb.2009.07.006 (2009).

24 Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458-472, doi:10.1016/j.cell.2012.01.010

S0092-8674(12)00016-5 [pii] (2012).

25 Lanctot, C., Cheutin, T., Cremer, M., Cavalli, G. & Cremer, T. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat Rev Genet* **8**, 104-115, doi:nrg2041 [pii]

10.1038/nrg2041 (2007).

26 Misteli, T. Self-organization in the genome. *Proc Natl Acad Sci U S A* **106**, 6885-6886, doi:10.1073/pnas.0902010106

0902010106 [pii] (2009).

27 Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R. & Mozziconacci, J. Normalization of a chromosomal contact map. *BMC Genomics* **13**, 436, doi:10.1186/1471-2164-13-436

1471-2164-13-436 [pii] (2012).

28 Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases

to characterize global chromosomal architecture. *Nat Genet* **43**, 1059-1065, doi:10.1038/ng.947
ng.947 [pii] (2011).

29 Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9**, 999-1003, doi:10.1038/nmeth.2148
nmeth.2148 [pii] (2012).

30 Alber, F. *et al.* Determining the architectures of macromolecular assemblies. *Nature* **450**, 683-694, doi:nature06404 [pii]
10.1038/nature06404 (2007).

31 Bau, D. & Marti-Renom, M. A. Structure determination of genomic domains by satisfaction of spatial restraints. *Chromosome Res* **19**, 25-35, doi:10.1007/s10577-010-9167-2 (2011).

32 Bau, D. *et al.* The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol* **18**, 107-114, doi:10.1038/nsmb.1936
nsmb.1936 [pii] (2011).

33 Rousseau, M., Fraser, J., Ferraiuolo, M. A., Dostie, J. & Blanchette, M. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *Bmc Bioinformatics* **12**, doi:Artn 414
Doi 10.1186/1471-2105-12-414 (2011).

34 Hu, M. *et al.* Bayesian inference of spatial organizations of chromosomes. *PLoS Comput Biol* **9**, e1002893, doi:10.1371/journal.pcbi.1002893
PCOMPBIOL-D-12-00762 [pii] (2013).

35 Tark-Dame, M., van Driel, R. & Heermann, D. W. Chromatin folding--from biology to polymer models and back. *J Cell Sci* **124**, 839-845, doi:10.1242/jcs.077628
124/6/839 [pii] (2011).

36 Tokuda, N., Terada, T. P. & Sasai, M. Dynamical modeling of three-dimensional genome organization in interphase budding yeast. *Biophys J* **102**, 296-304, doi:10.1016/j.bpj.2011.12.005
S0006-3495(11)05401-4 [pii] (2012).

37 Barbieri, M. *et al.* Complexity of chromatin folding is captured by the strings and binders switch model. *P Natl Acad Sci USA* **109**, 16173-16178, doi:DOI 10.1073/pnas.1204799109 (2012).

38 Grosberg, A. Protein folding in polymer physics context. *Nato Sci Ser I Life* **333**, 299-311 (2001).

39 Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113, doi:10.1038/nature11279
nature11279 [pii] (2012).

40 Wang, X. Q., Crutchley, J. L. & Dostie, J. Shaping the Genome with Non-Coding RNAs. *Curr Genomics* **12**, 307-321, doi:10.2174/138920211796429772
CG-12-307 [pii] (2011).

41 Starr, D. A. & Fridolfsson, H. N. Interactions Between Nuclei and the Cytoskeleton Are Mediated by SUN-KASH Nuclear-Envelope Bridges. *Annu Rev Cell Dev Bi* **26**, 421-444, doi:DOI 10.1146/annurev-cellbio-100109-104037 (2010).

42 Wood, A. J. *et al.* Targeted Genome Editing Across Species Using ZFNs and TALENs. *Science* **333**, 307-307, doi:DOI 10.1126/science.1207773 (2011).

43 Bickmore, Wendy A. & van Steensel, B. Genome Architecture: Domain Organization of Interphase Chromosomes. *Cell* **152**, 1270-1284, doi:10.1016/j.cell.2013.02.001 (2013).